QLUCORE

Guided variable filtering for PCA with the projection score

INTRODUCTION

In this overview document we will describe the basic characteristics of the projection score, a new and unique feature of Qlucore Omics Explorer that is aimed at providing guidance to the user for selecting an optimal variable subset via variance filtering.

We will use the term data set to describe the measured data. The data set consists of a number of samples for which a set of variables have been measured. All variables are measured for all samples.

WHAT IS THE PROJECTION SCORE?

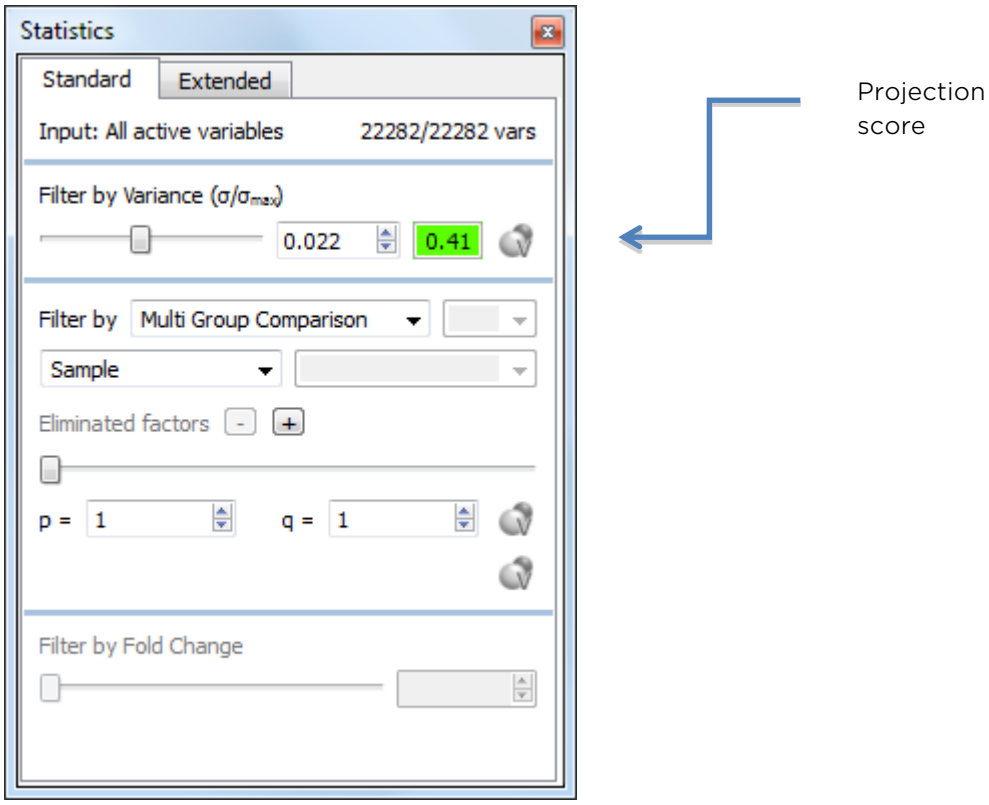
The projection score is a new tool that is unique to Qlucore Omics Explorer. It measures the informativeness of a low-dimensional representation obtained by PCA, and allows explicit comparison of representations corresponding to different variable subsets, e.g., those obtained by variance filtering of a large data set. The goal of exploratory visualization is to find a representation from which we can extract interpretable and potentially interesting information, that is, one that contains structures and patterns that are likely to be non-random. By following the evolution of the projection score in real time during variance filtering, the user can easily find the variable subset (and thus implicitly the variance cutoff) giving the most informative representation.

The information content of a PCA representation is often measured by means of the fraction of the total variance in the data set that is captured in the low-dimensional representation. This measure, however, is strongly dependent on the dimensions (the number of samples and variables) of the data set. Thus, it is not very useful for comparing the informativeness of representations of data sets with different dimensions, such as those resulting from variable filtering. As an example, consider two data sets with 3 and 10,000 variables, respectively. Regardless of the strength of the non-random signal in the small data set, the first three principal components will always capture 100% of the variance in the data (in other words, it will always contain a lot of information if we measure information by the captured variance fraction). In contrast, it is extremely unlikely that the first three principal components of the larger data set will capture even close to 100% of the variance, even if it contains strong, non-random and very interpretable patterns. From an interpretation point of view, it is thus clear that the fraction of captured variance alone is insufficient for evaluating the informativeness of a PCA representation. The idea behind the projection score is to quantify the informativeness not by the captured variance fraction itself, but by the excess of captured variance over that expected from a random data set (without any non-random, informative structure).

To compute the projection score for a given data set, we thus start by computing the fraction of the total variance that is captured by the first three principal components. Then, we estimate the expected value of the same entity for completely random data. The projection score is defined as the difference between the square root of the observed quantity and the square root of the expected value for random data. Hence, a large value of the projection score means that the PCA representation of the observed representation contains much more information (variance) than the corresponding representation of a random data set of the same size, which suggests that there are non-random, potentially interesting structures present in the representation. In contrast, a projection score close to zero indicates that the representation is not more informative than one of a random data set and that there are no broad, consistent patterns to be found by the PCA.

HOW TO USE THE PROJECTION SCORE

By monitoring the evolution of the projection score during variance filtering, the informativeness of the variable subsets thus obtained can be followed and the optimal one (the one with the highest projection score) can be found. In Qlucore Omics Explorer, the projection score is shown in a box next to the variance filtering slider in the Statistics dialog, and the box showing the projection score is colored according to the displayed value (see figure below). Red color indicates a low projection score, yellow color indicates a medium-high score and green color corresponds to a high projection score.



In practice, almost all real data sets contain some non-random structure, and therefore it is very uncommon to get a projection score very close to zero. The colors, and thus the boundaries between what is considered to be a "good" or a "bad" projection score, are based on our experience from applying the projection score to high-throughput (mainly gene expression microarray) data sets, and should be interpreted mainly as rough guidelines suggesting the quality of the representations. For other types of data sets, other quality cutoffs may be more suitable.

WHEN IS THE PROJECTION SCORE APPLICABLE?

In principle, the projection score is flexible enough to be used to compare variable subsets obtained from the original data set through any selection procedure. However, this can become computationally demanding since the same selection procedure has to be applied to a large collection of random data sets. For this reason, and to favor an instantaneous response that can be truly helpful in an exploratory setting, the projection score calculations in QOE are currently restricted to work in the context of variance filtering. In this context, the observed variance content can be compared to pregenerated values to obtain the projection score, which allows the values to be continuously updated alongside with the filtering.

Moreover, for consistency with the three-dimensional PCA plot, the current implementation computes the projection score based on the first three principal components.

WHERE CAN I GET MORE INFORMATION?

The rationale behind the projection score, as well as some examples of applying it to omics data, are more extensively described in the article

M Fontes and C Soneson: The projection score - an evaluation criterion for variable subset selection in PCA visualization. BMC Bioinformatics 12:307 (2011).

DISCLAIMER

The contents of this document are subject to revision without notice due to continuous progress in methodology, design, and manufacturing.

Qlucore shall have no liability for any error or damages of any kind resulting from the use of this document.

Qlucore Omics Explorer is only intended for research purposes.